(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 May 2002 (10.05.2002)

PCT

(10) International Publication Number
WO 02/37326 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/GB01/04869

(22) International Filing Date:
2 November 2001 (02.11.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0026936.5 3 November 2000 (03.11.2000) GB

(71) Applicant *(for all designated States except US)*: **ENVISIONAL TECHNOLOGY LIMITED** [GB/GB]; Westbrook Centre, Milton Road, Cambridge CB4 1YG (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SWANNACK, Christopher, Martyn** [GB/GB]; Envisional Limited, Westbrook Centre, Milton Road, Cambridge CB4 1YG (GB). **COPPIN, Benjamin, Kenneth** [GB/GB]; Envisional Limited, Westbrook Centre, Milton Road, Cambridge CB4 1YG (GB). **GRANT, Calum, Anders, McKay** [GB/GB]; Envisional Limited, Westbrook Centre, Milton Road, Cambridge CB4 1YG (GB). **CHARLTON, Christopher, Toby** [GB/GB]; Envisional Limited, Westbrook Centre, Milton Road, Cambridge CB4 1YG (GB).

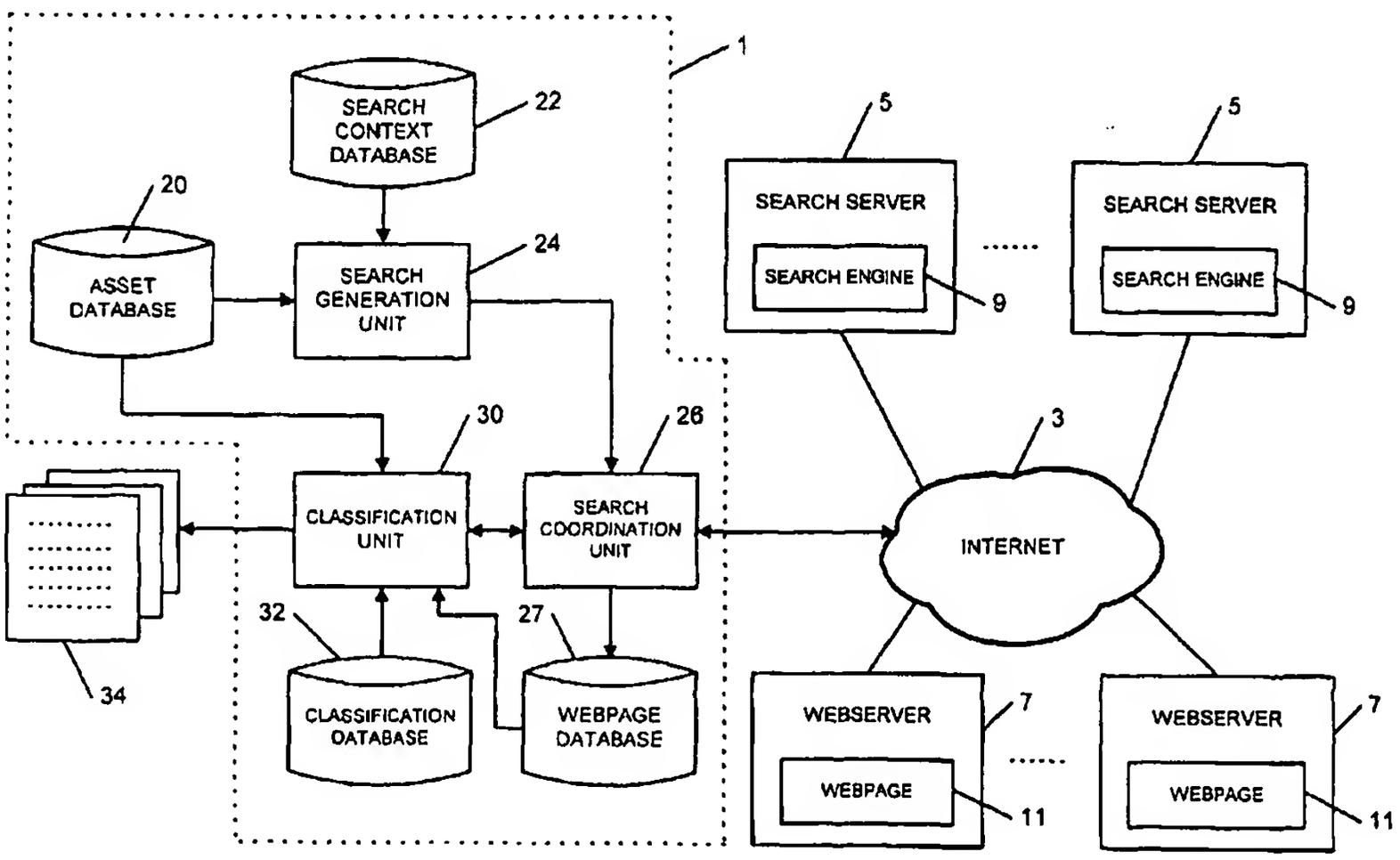
(74) Agents: **BERESFORD, Keith, Denis, Lewis** et al.; Beresford & Co., 2-5 Warwick Court, High Holborn, London WC1R 5DH (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,

[Continued on next page]

(54) Title: SYSTEM FOR MONITORING PUBLICATION OF CONTENT ON THE INTERNET



(57) Abstract: A monitoring system (1) is provided for identifying web page (11) accessible on the internet (3) relating to products identified by an asset database (20). The monitoring system (1) initially causes searches to be performed by search engines (9) for web pages (11) relating to the products. Retrieved web pages (11) are then classified by a classification unit (30). If a web page (11) is identified as being relevant more detailed analysis is performed to identify web pages (11) containing download links for digital files corresponding to the products. A report (34) identifying these web pages (11) is then generated.

WO 02/37326 A1

IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM FOR MONITORING PUBLICATION OF CONTENT ON THE INTERNET

The present invention is concerned with a system for monitoring publication of content on the internet. Embodiments of the present invention concern systems for monitoring for unauthorised publication of music on the internet and for monitoring the context in which companies and their products are discussed on the internet.

The development of the Internet has led to improvements in the ability to transfer information electronically from one computer to another. One consequence of this is that information is increasingly made available on computer databases for electronic retrieval. This means that more information is now disseminated to a wider audience, about a more extensive range of subjects.

Information about commercial activities, much of it unofficial and possibly commercially damaging, can be disseminated to consumers with ease. Commercial operations are sensitive to the publication of this type of information, because it can have deleterious effect on the reputation of the business. For example, false

information about the efficacy of pharmaceuticals or safety of foodstuffs can be circulated to a wide audience, before a commercial entity becomes aware of the information. By the time the commercial entity has managed to take steps to prevent the further circulation of information, that information may already have had a commercially damaging effect.

In addition to text published on the internet any other type of digital recording may also be distributed. Where the digital recordings are unauthorised copies of copyright works, it can be important to identify unauthorised distribution so that it can be stopped.

Items of information concerning a particular subject for retrieval via the Internet can be sought and identified by means of search engines. Most search engines are operable to receive an input consisting of a string of text. This string of text is known as a search string, which is used by the search engine to find matches, or near matches, in the content of items of information accessible to the search engine. Such items of information can include websites and newsgroups. The search engine then presents a list of results to the user. The list identifies websites and newsgroups

considered by the search engine to have a match with the search string. The match can be an exact match, or provision can be made for the search engine to identify near matches to the search string, near matches being determined by truncations, letter transpositions or letter replacements within the search string.

A disadvantage of the search engine of this type is that it can deliver erroneous results. For example, if the search string is too short, or relates to too general a subject, then a match to the string may be found in a large number of websites. The content of many of those websites may be wholly unrelated to the subject matter of the search string, the inclusion of the search string in the website being entirely coincidental. Thus, if an investigator making use of a search engine on behalf of a commercial entity searches on the basis of a well known trade mark, many instances of use of that trade mark may arise which are of no interest to the investigator. Review of all of these websites can be labourious and extremely time consuming.

Also, many search engines make use of "meta tags" which are strings of text embedded in web page descriptions by a web page designer but which do not cause display

output. Meta tags are used by web designers to maximise the chance that a website will be identified by a search engine as relating to a particular subject. However, it may be commercially advantageous to a web designer to include a large number of meta tags relating to diverse subjects, causing a search engine to erroneously identify a website as relating to a search string not entirely related to the subject matter of the website, so that the website is regularly found by search engines and thus receives more commercial exposure. An investigator can find this disadvantageous because many websites may be identified with a search engine, which include meta tags which relate to the search string, but which are in fact not relevant to the subject matter defined by the search string.

On the other hand, if the investigator chooses a search string which is too long or too specific, investigation may not be sufficiently thorough, because many websites may be overlooked by the search engine which, in fact, relate wholly to the subject matter of the search string but which do not contain text which exactly or nearly exactly matches the search string.

Furthermore, some search engines provide collated

information to a user. This information consists of identified websites and newsgroups, categorised by subject matter. These categories are presented to the user in a hierarchical tree structure; the category headings can be searched with respect to a search string in the same way as described above in relation to a search of website contents. However, a disadvantage of this arrangement is that it relies on the investigator understanding the manner in which websites have been categorised into particular categories in the hierarchical structure, and for the investigator to check the correct categories for the subject under investigation. It is possible that the investigator might overlook categories which are of relevance, or that the person who categorised the websites into the categories might have wrongly categorised a website into a category which the investigator does not consider sufficiently relevant as to warrant investigation. This can mean that an investigator can overlook websites which are of relevance to the subject under investigation. Also a website investigator might find checking a large number of categories, to ensure the thoroughness of the search, labourious and time consuming.

In addition to performing searches using search engines,

an investigator working on behalf of a commercial organisation to establish whether that organisation is being discussed in a potential commercially damaging manner, can make investigations of messages being posted in newsgroups. Newsgroups are facilities operable using network news transfer protocol (NNTP) which allow messages to be posted in a central server for retrieval and review by users. The contents of newsgroups can be highly dynamic, with the contents of a newsgroup typically being replaced every three days. Thus, for an investigator to monitor the contents of newsgroups can be time consuming and labourious. A large number of newsgroups and a large number of messages on each newsgroup must be reviewed in order to establish whether any damaging messages are being posted. Also, if an investigator finds it necessary to check a large number of newsgroups, it may not be possible to review all messages in the time available before messages are deleted from the newsgroup and new messages are posted.

Whereas search engines are configured to search and identify newsgroups as relating to a subject signified by a search string, they generally only search newsgroup headings, and newsgroup descriptions if available. Messages posted on newsgroups may contain relevant

information, but will not be detected since the search engine will not search through messages.

Therefore, it is an object of the invention to provide a system capable of collecting data and processing the data to present relevant data therein to a user.

It is a further object of the invention to provide a system capable of configuring search engines to retrieve and classify data in accordance with a user requirement.

It is another object of the invention to provide a system operable to monitor published data sources for relevant information and to deliver relevant information as required.

These and other objects may be achieved, wholly or in part, by the invention, aspects of which are set out below.

One aspect of the invention provides means for storing instructions for transmittal to a search engine for generation of search results, means for receiving search results retrieved by a search engine in response to one of said instructions, and means for processing said

search results to establish which of said results are sufficiently relevant, relative to a user determined relevance criterion, to be output to a user.

Another aspect of the invention provides means for storing instructions for transmittal to search engines, means for retrieving search results from search engines in response to said instructions, means for retrieving, in accordance with said search results, items of information corresponding to said search results, and means for processing said items of information to identify relevance or otherwise thereof.

Another aspect of the invention provides apparatus for retrieving and processing information comprising means for storing instructions for retrieval of information, means for storing retrieved units of information and means for identifying relevance of said information in accordance with predetermined criteria.

In accordance with another aspect of the invention, apparatus is provided which comprises means for receiving a user input instruction indicating a document relevance criterion, means for reviewing the content of an item of information with respect to said received instruction,

and means for storing a value representative of the relevance of said item of information with respect to said document relevance criterion.

Another aspect of the invention provides apparatus for retrieving and processing information held in units in a remote location, comprising means for retrieving information in accordance with a predetermined sequence, and discrimination means operable to test a unit of retrieved information against one or more predetermined criteria and to generate a score for said unit of information on the basis of said one or more criteria.

Further aspects and advantages of the invention may become apparent from the following description of specific embodiments of the invention, with reference to the accompanying drawings in which:

Figure 1 is a schematic diagram of a network of computers connected via the internet, including an internet monitoring system in accordance with a first embodiment of the present invention;

Figure 2 is a flow diagram of the overall processing of the monitoring system of Figure 1;

Figure 3 is a flow diagram of the processing of the monitoring system of Figure 1 to initiate a search;

Figure 4 is a flow diagram of the processing of the monitoring system of Figure 1 to expand a search;

Figure 5 is a flow diagram of the processing of the monitoring system of Figure 1 to identify music piracy websites;

Figure 6 is a schematic diagram of a network of computers connected via the internet including an internet monitoring system in accordance with a second embodiment of the present invention; and

Figure 7 is an illustration of an exemplary user interface of the monitoring system of Figure 6.

FIRST EMBODIMENT

Figure 1 is schematic block diagram of a computer network comprising a monitoring system 1 connected via the internet 3 to a plurality of search servers 5 and a plurality of web servers 7. In this embodiment, the monitoring system 1 is arranged to utilise search engines 9 available on the search servers 5 to identify web pages

11 posted on the web servers 7 which relate to the sale and distribution of unauthorised copies of music tracks.

In order to identify web pages 11 relating to the unauthorised distribution of identified music tracks, the monitoring system 1 of this embodiment comprises an asset database 20 identifying the artists and songs, associated with tracks, unauthorised distribution of which is to be monitored; a search context database 22 and a search generation unit 24 to process data on the asset database 20 and utilise that data to generate seed terms for initiating the search via the internet 3 for web pages 11 relating to unauthorised distribution of the assets identified by the asset database 20; a search coordination unit 26 for scheduling searches to be performed by the search engines 9 and expanding an initial search utilising links identified in web pages 11 located by the search engines 9; a web page database 28 for storing web pages 11 retrieved as a result of the searches coordinated by the search coordination unit 26; a classification unit 30 arranged to retrieve data from the asset database 20 and a classification database 32 and for processing retrieved web pages 11 stored in the web page database 28 to assess the relevance of web pages 11 and output a report 34 identifying web pages

distributing unauthorised copies of the recordings being monitored.

The monitoring system 1 in accordance with this embodiment of the present invention is arranged to perform a search of web pages 11 available via the internet 3 that is simultaneously extensive and focussed. The monitoring system 1 achieves these contradictory aims by initially initiating a large number of searches performed by the search engines 9 on the search servers 5 which all relate to the songs and artists identified by the asset database 20 together with additional data identifying the context (in this example unauthorised distribution of recordings) identified by the data stored in the search context database 22. The web pages 11 identified in these searches are then retrieved by the search coordination unit 26 via the internet 3 and stored in the web page database 28. Each of the web pages 11 will comprise HTML (Hyper Text Mark Up Language) data identifying the text and layout of the web pages 11. The classification unit 30 then processes the HTML of the retrieved web pages 11 using data from within the asset database 20 and classification database 32 to form an initial assessment of the relevance of the retrieved web pages 11. If the retrieved web pages 11 are identified

as being relevant, further web pages 11 identified by links within the relevant web pages 11 are utilised by the search coordination unit 26 to retrieve further web pages 11 via the internet 3 for storage within the web page database 28. When sufficient web pages 11 have been stored within the web page database 28 the web pages 11 identified as being most relevant are then analysed in greater detail by the classification unit 30 so that a report 34 identifying the most relevant pages can be generated.

All searches are necessarily a compromise between extent of recall and precision of identification of documents. For the general public, the primary criteria for a search engine is precision. That is to say for most searches, it is desirable that only documents relevant to the search criteria are identified by a search engine 9. To that end, commercial search engines 9 are tuned to identify fewer more relevant documents thus increasing the chances that documents identified will be relevant at the expense of missing documents.

The applicants have appreciated that for a monitoring service identifying data accessible on the internet, it is only necessary to identify documents which can be

found rather than all documents of interest which are in existence. Thus by performing query expansion and submitting multiple search requests to a large number of commercial search engines, the majority of accessible web pages of potential interest can be identified. Once web pages of potential interest have been identified a much more detailed analysis of retrieved web pages can be performed. As the criteria of interest are known, it is possible to define the manner by which documents are to be analysed to determine whether they are of interest in detail in advance. Thus for the limited aim of monitoring documents available on the internet in relation to a highly specific criteria, reports can be automatically prepared.

The processing of the monitoring system 1 of Figure 1 will now be described in detail with reference to Figures 2-5.

Figure 2 is an overview flow diagram of the processing of the monitoring system 1 of Figure 1. Figures 3-5 are each detailed flow diagrams of processing taking place at steps illustrated in Figure 2.

Initially when the monitoring system 1 is activated to

monitor web pages 11 accessible via the internet 3, this causes the search generation unit 24 to access the asset database 20 and search context database 22 to generate a series of search queries which are submitted by the search coordination unit 26 to the search engines 9 on the search servers 5 accessible via the internet 3 (S2-1).

Referring to Figure 3 initially (S3-1) the search generation unit 24 accesses the asset database 20 to obtain a list of assets which are to be monitored. In this example, the asset database 20 is taken to comprise a list of songs and artists identified assets owned by a music company. For each song and artist combination, the search generation unit 24 then proceeds to process the artist/song title combination to remove punctuation from the text retrieved from the asset database 20. The search generation unit 24 then (S3-2) filters the text from which punctuation has been removed, deleting stop words from the retrieved text which have a very high frequency in the English language and hence are unsuitable for being used as a search term to locate web pages 11 related to the song and artist.

Thus for example assuming the asset database contained

only a single record of the following form:

Title: Oops! I did it again

Artist: Spears, Britney

the search generation unit 24 would then process the record obtained from the asset database 20 removing punctuation and capitalising all words to generate text data of the following form:

Text Data: OOPS I DID IT AGAIN SPEARS BRITNEY

the search generation unit 24 then (S3-2) deletes from the text data words having a very high frequency in the English language which in the above example would be the words I, did, it and again to obtain search data of the following form:

Search Data: OOPS BRITNEY SPEARS

the search generation unit 24 then proceeds to retrieve (S3-4) from the search context database 22 additional key words to be used to initiate searches on the search engines 9 accessible via the internet 3. The search context database 22 is arranged to contain a list of key

words identifying the context in which discussion of the asset identified by the records with the asset database 20 are discussed in web pages 11 accessible via the internet 3. In this embodiment which the monitoring system 1 is arranged to monitor unauthorised distribution of recordings via the internet, the context words stored on the search database 22 would be context words related to the downloading of free music from web pages 11 via the internet 3 and could be for example:

Context Words	:	FREE MUSIC
	:	DOWNLOAD
	:	MP3
	:	etc.

Once a list of context words has been obtained from the search context database 22, the search generation unit 24 then generates a list of searches which are to be initiated by the search coordination unit 26. This list of searches comprises all combinations of the search data generated from a record retrieved from the search database 20 together with each of the context words retrieved from the search context database 22.

Thus for example in the case of the search data and context words mentioned above searches including the following would be created and passed to the search coordination unit 26:

Searches: OOPS
: BRITNEY
: SPEARS
: OOPS & BRITNEY
: OOPS & SPEARS
: BRITNEY & SPEARS & OOPS
: BRITNEY & FREEMUSIC
: etc.

When a list of searches has been passed to the search coordination 26 the search coordination unit 26 then proceeds to schedule (S3-5) search requests to be passed to the search engines corresponding the searches received from the search generation unit 24. In this embodiment, the search coordination unit 26 is such to interrogate a large number of commercially available search engines 9 such as those run by Yahoo!, Google, Hotbot, Lycos, Alta Vista, etc. In order to avoid overburdening an individual search engine 9, the search coordination unit 26 in this embodiment is arranged to schedule the

searches to be performed so that queries are dispatched to individual search engines 9 at a rate set by a user up to a frequency of 1 every 5 seconds. In other embodiments interrogation of search engines 9 may be arranged to occur at specific times or on specific days so that the processing by the search engines 9 is appropriately scheduled.

Returning to Figure 2 once the searches scheduled by the search coordination unit 26 have been initiated (S2-1) the search coordination unit 26 then proceeds to wait until HTML pages are received from the various search engines 9 which have been asked to process the initiated searches. The search coordination unit 26 then (S2-2) proceeds to classify received search results and expand the extent of the monitored search space accordingly which will now be described in detail with reference to Figure 4.

Initially (S4-1) the search coordination unit 26 waits until search results are received from the search engines 9 to which queries have been submitted. When a search result is received, the search results comprise HTML scripts containing links to web pages identified as being relevant by the search engines 9. The search

coordination unit 26 then (S4-2) proceeds to extract from the HTML scripts received from the search engines 9 each of the HTML links and stores the links not previously stored within the web page database 28.

By generating searches in the manner described and submitting search queries to a large number of commercially available search engines 9 a database of HTML links containing no duplicates is stored within the web page database 28.

The applicants have appreciated that although this list of links stored within the web page database 28 is not a complete list of web pages 11 relating to assets identified by the asset database 20, this list will contain virtually all links which an individual attempting to obtain a pirate copy of the recording identified by the asset database 20 might identify. The fact that further unidentified web pages 11 may exist is not of significant concern as the monitoring system 1 is intended to monitor web pages 11 that are likely to be accessed rather than all possible web pages.

Once a list of links has been placed within the web pages database 28 the search coordination unit 26 then

proceeds to download (S4-3) each of the web pages 7 identified by the links stored within the web page database 28. In practice, as the queries submitted to the various search engines 9 are scheduled by the search coordination unit 26 at different times, the downloading of individual web pages 11 and the storage and retrieval of links to be placed in the web page database 28 happen simultaneously but have been described here as happening consecutively in the interests of clarity and brevity.

Once HTML for the next unprocessed link has been downloaded by the search coordination unit 26 (S4-3) the search coordination unit 26 passes the retrieved HTML to the classification unit 30. The classification unit 30 then proceeds to determine a set of classification scores (S4-4) for the retrieved HTML data.

When processing a HTML for a retrieved web page 11, initially, a preliminary check is carried out by the classification unit 30 against one or more definitions stored in the classification database 32. Each of these definitions is a collection of words, each assigned weightings corresponding to expected frequency in a piece of text. A word having a low frequency in an average piece of text is assigned a high weighting, and a word

having a high frequency in an average document is assigned a low weighting. In certain cases, such as for example in the case of the word "the", words are assigned a zero weighting.

The actual incidence of words in the submitted data is tested against the words contained within the definitions and a collective score is obtained for the definition in relation to the submitted data. This weighting gives a general impression as to the relevance of the data to a particular definition. In order to compensate for the possible grammatical inflection of words in a piece of data, a stemming function may be applied to the words in the definition. In the present embodiment, such stemming is optional. Following this initial relevance check, the retrieved HTML document is then processed against a detailed set of rules stored within the classification database 32 for a more thorough classification of the data.

In this embodiment, this is achieved by the classification unit 30 analysing the contents of the retrieved HTML and compiling scores for each web page. This is done by applying rules stored within the classification database 32. The classification unit is

able to analyse rules according to a rules definition language which provides a user defining a rule with a facility to match words exactly, with case sensitivity, according to similarity, according to phonetic match, a semantic match and a stemmed match. Also the rules language allows rules to be established which test for the distance between words, the position of the word in the document, for example by means of paragraph number or sentence number or location (title, authorship or heading). The formulation and processing of HTML in accordance with rules will be described later.

The result of the classification according to the rules is a list of categories and scores for the retrieved HTML document. The classification unit 30 manages different categories of scores for a document and returns a list of categories and scores for that document once the review of the document has been completed. Scores can be calculated (depending upon the manner in which rules are programmed by a user) on the basis of different scoring methods.

For example, a cumulative scoring allows a score to be added each time a condition is met in a document, a one off scoring basis allows a score to be added to a

category only once for a particular document (so that later instances of a particular condition having been met have no impact on the score), or in a weighted basis. A weighted basis is exemplified by an exponential decay, whereby a score is added to a total score for a document on each occasion that a condition is met, with the additional score becoming repeatedly smaller on each additional occasion that the condition is met. Positive and negative weightings can be provided.

Returning to Figure 4, once the classification unit 30 has generated a set of classification scores for a particular retrieved HTML web page 11, the results of the classification are stored with the HTML data within the web page database 28. The search coordination unit 26 then determines (S4-5) whether any of the classification scores for a retrieved web page is sufficiently high indicating that the retrieved web page is considered of relevance to the monitoring being carried out.

If this is determined to be the case the search coordination unit 26 then proceeds to add to the links stored within the web page database 28 any HTML links included within the classified web page 11 provided that the links are not already stored within the web page

database 28 and provided that the links encountered were not at a depth greater than the maximum value for expanding search. In this embodiment, this is achieved by storing with each of links identified by the search engines a depth of value of zero and incrementing the depth value stored with the link each time further links are retrieved from a web page classified as being relevant.

Thus in this way in addition to all of the web pages initially identified by the search engines 9, the web page database 28 also has stored within it additional web pages identified by following links through the initial pages identified by the search engines 9 where the pages identified by the search engines 9 have been classified by the classification unit 30 as being relevant to the monitoring being performed.

It will be appreciated that whereas in this embodiment all links on a page are stored in the web page database 28, in other embodiments links might be selectively stored. Thus for example whether links are stored could depend upon text associated with an artist appearing on or near a link. Alternatively where links are associated with certain kinds of text e.g. an alphabetic list they

might automatically be selected for inclusion in the web page database 28. The rules utilised to select links could be dependent upon the classification scores for the web page containing the links with different rules applying to different classifications.

After any additional links have been added to the web page database 28 (S4-6) the search coordination unit 26 then determines (S4-7) whether HTML data has been retrieved for all the links stored within the web page database 28. If this is not the case the next link within the web page database 28 is then utilised to retrieve HTML (S4-3) which is classified and further links are obtained if that classification is considered relevant to the search being performed (S4-4 - S4-6).

Returning to Figure 2, eventually when all of the HTML pages identified by links within the web page database 28 have been retrieved and classified by the classification unit 30, the classification unit 30 then proceeds to process the HTML pages considered to be most suspect as identified by their classification scores in detail (S2-3) which will now be described with reference to Figure 5.

Referring to Figure 5, the classification unit 30 initially (S5-1) identifies all of the web pages 11 within the web page database 28, associated with classification scores indicating a high level of relevance for the monitoring being performed. In this embodiment, where the classification database 32 is arranged so as to identify web pages 11 relating to the downloading of music to be given high classification scores, it is these web pages which are considered for further detailed analysis.

Once the web pages 11 in the web page database 28 associated with high classification scores have been identified, the classification unit 30 then selects (S5-2) the first of the web pages 11 associated with a high classification score for further analysis. The classification unit 30 then processes the selected web page by identifying the location of key words in the HTML text and their screen position relative to music download links appearing in the page.

Specifically, in this embodiment of the present invention the classification unit 30 initially processes records within the asset database 20 to generate a list of key words corresponding to the search data obtained by the

search generation unit 24 when processing the asset database 20. That is to say the classification unit 30 obtains from the asset database a list of key words being uncommon words appearing in song titles and artist names of the records within the asset database 20. The classification unit 30 then processes the selected web page to identify whether any of these key words appear in the currently selected web page.

Additionally, music download links in the retrieved pages are also identified. In this embodiment, the links within an HTML page for downloading music are identified by the classification unit 30 noting HTML links referencing files with one of a specified number of identifiers such as ".mp3" ".m3u" or ".zip" which are identified as being linked to files of at least a certain size indicative of the link being a downloading link for a music recording. Data identifying the position on an output page where the link identified as being a music download link is then stored.

In this embodiment the size of an individual file for downloading is determined by the monitoring system 1 initially attempting to download the identified file. Conventionally, this causes data identifying the size of

the file located by a hyperlink to be returned. Once data identifying the size of a file has been determined the download operation is then aborted.

If (S5-4) the currently selected web page is determined to have both a music download link and one or more of the key words obtained from the asset database 20 the classification unit 30 then (S5-5) processes the HTML for the selected web page to identify whether any of the other words appearing in either the artist name of song title including the key word appears in the vicinity of the identified key word.

Thus for example if given the asset database previously described, a web page identified as including a link to an MP3 file or the like to a large data file is found to indicate the word "Oops" displayed close to the link, the classification unit 30 would then search words adjacent to "Oops" to identify whether any of the other words associated with either "Britney Spears" or "Oops I did it again" were also included in text which would be displayed in the portion of the screen near to the link to the MP3 file. The classification unit 30 then generates a score for the web page based upon the extent to which the web page includes data corresponding to the

record adjacent to the music download link. In this embodiment, the classification unit 30 is set to identify a web page as a music download link for an identified asset if 60% of the words associated with an asset appear in the web page near to the music download link. If this is the case, the classification unit notes the web page as a potential music piracy page and then (S5-6) goes on to determine whether all of the web pages identified as relevant have been processed. The requirement that 60% of the words appear increases the likelihood that a download associated with a musician/song title is identified whilst allowing for variations in the way in which a title/artist are written on the suspect page. If this is not the case the next page is then selected and processed (S5-2 - S5-5).

It will be appreciated that for other embodiments different variations on the amount of text differs from data within the asset database 20 could be permitted. Thus for example for monitoring for films or computer software where only the name of a product is likely to appear, a matching requirement of 90% might be selected. Alternatively different levels match might be applied to for example the name of a song on a page (low match) and the name of a song on a page determined to contain an

artists name (higher match to song title required).

Returning to Figure 2, finally, after all the relevant web pages in the web page database 28 have been identified and processed the classification unit 30 then outputs (S2-4) a report on the suspect sites, being those sites which are considered relevant and those where music download links have been identified. This report could be of the form of a list of the web pages identified as relevant within the web page database 28 together with complete screen dumps generated with the HTML scripts for the web pages identified as being related to specific music downloads.

In this embodiment, in order to determine whether a web page 11 is retrieved and hence to be considered for detailed analysis the classification unit 30 processes HTML for web pages 11 stored on the web page database 27 against rules stored in the classification database 32. This classification is achieved by the classification unit parsing the HTML to identify words or phrases which match portions of rules defining how scores are to be generated in the event of a match.

Further features of the rules language will now be

described. The rules language is defined by the function of the parser in its ability to recognise functional words or phrases in a string of text.

Specifically, in the present embodiment rules are defined by text data which is then parsed to identify functions which generate scores based upon the appearance of text data within an HTML script. Example basic functions are explained in greater detail in the appendix. More complex rules to produce classification scores specific to an individual defined subject are then created from the basic functions as will now be explained.

Firstly, the rules language allows for words in a document to be matched to produce classification scores.

For example, the rule

for "dog" classify Canine

states that every time the word "dog" is encountered, the score for the classification "Canine" is incremented. At the beginning of the document, the score for Canine is set to zero. Basic word matching is not case sensitive.

More rules can be added, such as

```
for "cat" classify Feline
for "dog" classify Animal
for "cat" classify Animal
```

Note that in this example, the same word can be matched more than once, and that the same class can be matched more than once. Statements can be combined in curly brackets, so that the above rules could be rewritten

```
for  "cat"
{
     classify Feline
     classify Animal
}
for  "dog"
{
     classify Canine
     classify Animal
}
```

These rules return scores for three classes: Feline, Canine and Animal.

In addition to exact word matching described above, one of a list of words can be matched using the "or" operator. For example

```
for "computer" or "software" or "program" classify
    Computers
```

would increment the score for Computers each time one of the words in the list was found. This is equivalent to writing the three rules

```
for "computer" classify Computers
for "software" classify Computers
for "program" classify Computers
```

Combination of words can also be matched, by combining them with the "and" operator. For example

```
for "Bill" and "Gates" classify Microsoft
```

must find both the words "Bill" and "Gates" to call the classify statement. "and" and "or" can be used at the same time, so that

```
for "Bill" or "William" and "Gates"
```

matches either "Bill" or "William" and the word "Gates". Note that, in this rules language, the "or" operator has higher precedence than the "and" operator, which is contrary to normal operator precedence.

A stemming algorithm can be applied which stems each word before it is looked up. The keyword "stemmed" is inserted before the word to indicate that any stem of the word can be matched

for stemmed "pony" or stemmed "horse"

matches any stemmed word including "ponies" and "horses". A phonetic match can be made by inserting the "sound" keyword in front of the word. The rule:

for sound "Clinton" and sound "Lewinsky"

is likely to be able to match misspellings of the names "Clinton" and "Lewinsky". A case sensitive match can be specified by the "name" keyword. In this case,

for name "Clinton"

only matches the word Clinton if an instance of the word

in a document matches the word exactly, including taking account of upper case letters. Phrases can also be matched, so that

```
for  name "Bill Clinton"
for  stemmed "fish cake"
```

does a case sensitive match for the phrase "Bill Clinton" and a stemmed match for the phrase "fish cake".

Words, links and images an also be matched. This counts the number of words, links and images in the document:

```
for  word classify Word
for  image classify Image
for  link classify Link
for  "Michael Douglas" and image
          if near (1, 2) classify MichaelDouglasPicture
```

The last rules only matches if the phrase "Michael Douglas" occurs near an image.

The basic "classify" statement increments the class score by one. To adjust the class score by a different number, a weighting can be specified. This example adds 40 to

the score for English each time the word "the" is encountered. This rule is formulated because the word "the" is highly associated with the English language, and so can be used to give a high level of assurance that the document is in English.

```
for  "the" classify English weight 40
```

A negative weighting can be given, such as

```
for  "le" {classify English weight -3 classify French
     weight 2}
```

An arbitrary expression can be used to specify the weighting, such as

```
for  "hen" classify Poultry weight 2 * x - square (4)
```

By convention, there is a class name called "this" which is a class score for the agent currently being prepared. So the rule

```
for  "Madonna" classify this
```

would add one to the "this" score. Rules can also be

"accepting" or "rejecting", which add large positive or negative numbers to the class score. The following rules reject the class Currency if the word "stirling" is found, but accept the word "sterling" is found.

```
for  "stirling" reject Currency
for  "sterling" accept Currency
```

A rule can also set the weight of a score. For example

```
for  "jeans" classify Music set 0
for  "jeans" classify Clothing set 20
```

A classification can be adjusted just once, so that

```
for  "the" classify English weight 15 once
```

would increase the score for English by 15 only once. The maximum number of times a rule is invoked is specified

```
for  "the" classify English weight 10 max 4
```

which limits the contribution of this rule to 40 points. The contribution each weight makes to the score can be

made to decrease exponentially. The following example adds a maximum of 80 points to the class "Computers."

```
for  "program" classify Computer weight 80 exp
```

The first time the word "program" is reached, 40 is added to the Computer class score. The scores 20, 10, 5, 2, 1, ... are added as subsequent matches are found.

The rules language also allows for conditions to be included in rules. Conditions allow classification statements to be executed conditionally. Conditions can appear inside or outside "for" statements. A condition appearing inside a "for" statement can test for the relative positions and locations of the matched words. For example

```
for  "Bill" and "Gates" if near (1,2) classify Microsoft
```

classifies Microsoft if the first word is near the second word. An "else" clause can be given, so that

```
for  "Bill" and "Gates"
     if near (1, 2) {accept Microsoft classify Legal}
     else           classify Microsoft weight 3
```

"If" statements can be nested. Other textual conditions can be tested, and are listed in an appendix hereto. For example, the word position, sentence number, paragraph number, section number, and distances can be evaluated. The location can be tested to see whether it appears in a meta-tag, a link, a heading, or the title or if it is in bold, italic or is underlined.

A condition appearing outside a "for" statement can test general conditions about the document and query the class scores.

```
for  "der" or "das" classify German
if German
{
     for "Berlin" or "Heidelberg" classify GermanTourist
}
else
{
     for "the" or "it" classify English
     for "le" or "la" classify French
}
```

A score for a class is only updated after the classify statement that set it. Therefore a condition that tests

the value of a class must occur in the text after classify statements that update the score.

A condition is taken to be true if it evaluates to a positive number. If the value is zero or negative, the condition is false.

Many functions such as "near", "distance", "position", "sentence", and "paragraph" accept word numbers as their arguments. Every "for" statement must match a list of phrases, and the word number is its position in the "for" statement. The following rule is matched if the first phrase ("Uma Thurman") is near the second phrase ("Nike Trainers")

```
for  name "Uma Thurman" and "Nike Trainers"
     if near (1, 2) // . . .
```

The following rule is matched if either "Bill Gates" or "William Gates" (the first phrase) occurs in the same sentence as "richest" or "wealthiest" (the second phrase).

```
for  "Bill Gates" or "William Gates" and
     stemmed "richest" or stemmed "wealthiest"
```

```
if sentence (1) = sentence (2)      // ...
```

The following example must match 3 different phrases, and tests to make sure that they all appear in the same section of a document.

```
for  "Bill Gates" and
     "Judge Jackson" or "Jackson" and
     "breakup" or "split"
     if section (1) = sentence (2) = section (3) //...
```

Every expression in the described rule language has a fixed point floating point type. Booleans are represented as true = 1.0 and false = 0.0. Each string is translated to an integer index, which is similar to a pointer as used in C.

Function calls have the general form

```
function_name (arg1, arg2, ...)
```

where "function_name" is the name of a built in function, and arg1, arg2 ... are themselves expressions. The statement

```
print("Invalid input\n")
```

calls the "print" function to output the given string. Note that escape characters may be used in the string. Each function must receive the correct number of arguments, or a compile-time error occurs. Each function also has a numerical return value, so in this example the links () function returns the number of links in the page

```
if links ()> 20 accepts linkspage
```

The name of a class evaluates to its score, so that the expression

```
German > 30
```

evaluates to true if the class score for German is greater than 30.

It should be noted that expressions can be evaluated in two different circumstances. The first circumstance is when a word has been matched, so is before the entire document has been processed. These expressions occur within a "for" statement. In this case, the class scores are all zero, and some functions such as links () and

images () return incomplete results. Expressions that are executed outside "for" statements are executed after the whole document has been processed, and the class values can be used.

The comparison operators = (equal to), != (not equal to), < (greater than), >= (less than), >= (greater than or equal to) and <= (less than or equal to), return the Boolean value 0 or 1 depending upon the comparative values of their operands. "Not", "and" and "or" are fuzzy Boolean operators, described in the next section.

The standard arithmetic operators in the language are available, including + (addition), - (subtraction), * (multiplication), / (division) and % (modulo). Normal operator precedence applies, and round brackets can be used to group expressions.

All truth values in the rule language are fuzzy, and are represented as continuous belief values within the range 0 to 1 inclusive. For example a degree of belief of 0.2 represents a relatively unlikely circumstance, while 0.99 represents a highly likely circumstance. In fuzzy logic,

- not x evaluates 1 - X

- x and y evaluates to the minimum of X and Y
- x or Y evaluates to the maximum of X and Y

The statements

```
P(Burglary)     =     0.001
P(Earthquake)   =     0.002
```

assign the values 0.001 and 0.002 to Burglary and Earthquake respectively, and is equivalent to

```
Burglary        =     0.001
Earthquake      =     0.002
```

Conditional probabilities are expressed as

```
P(Alarm | Burglary and Earthquake)              =     0.95
P(Alarm | Burglary and not Earthquake)          =     0.95
P(Alarm | not Burglary and Earthquake)          =     0.95
P(Alarm | not Burglary and not Earthquake)      =     0.95

P(JohnCalls | Alarm)                            =     0.95
P(JohnCalls | Not Alarm)                        =     0.05
```

P(MaryCalls | Alarm) = 0.70

P(MaryCalls | not Alarm) = 0.01

The probabilities form a belief network that can propagate values forwards through the network. The above example calculated probabilities (or belief values) for Alarm, JohnCalls and MaryCalls given the initial conditions Burglary and Earthquake. Changing the initial conditions (for example as a result of document analysis) propagates different belief values through the network.

The result is a set of probabilities (or belief values) for various properties about the document.

Further features of the rules language are now set out below.

Comments

Comments are written in C++ style, for example

```
for  "cat" and "mouse"   // Matches cartoon characters
```

In this example, the comment is "Matches cartoon characters". The text of the comment is purely for guidance of the human operator and this text is

disregarded by the parser.

Compound Statements

A statement may be composed of a list of other statements, in curly brackets. For example

```
for  "der" or "das" and "kapital"
{
     classify German
     if near (1, 2) accept Book
}
```

While Loops

While loops are executed while the condition is true. The following example sums the first 10 integers.

```
x = 10
y = 0
while X > 0
{
     y = y + x
     x = X - 1
}
```

Function Calls

A function call can also be used as a statement, for example

```
if links () > 100
     print ("This looks a bit like a links page.\n")
```

Assignment Statements

An alternative notation for

```
classify x set x + 1
```

is

```
x = x + 1
```

The following example computes the factorial of 10.

```
x = 10
Factorial = 1
while x > 0
{
     Factorial = Factorial * x
     x = x - 1
}
```

Return Statements

A class can be tagged as "returned" meaning that the

class value should be treated as a return value. This does not affect the running of the rules. The following example tags "English", "French" and "German" as valid return classes - other classes are ignored.

```
return English
return French
return German
```

In addition to generating classification scores by processing rules scores in this embodiment are also generated by comparing word frequencies in documents against prestored word frequencies for different document types. The prestored word frequencies for documents of different types can be calculated directly by processing known documents of a particular type. Additionally, information about the content of a document may be inferred from the frequency of the use of words in that document. For such a system documents relating to a particular subject can be processed to establish usual frequencies with which words appear in documents related to different subjects.

The rules language and processing of documents in addition to identifying relevant documents may also be

arranged to filter certain documents from consideration automatically. This can be achieved by the rule language identifying the presence of certain phrases or words as indicating a document not to be of interest and hence not to be processed further.

Thus for example a rule might be provided to exclude from monitoring a list of known authorised or reputable websites or alternatively certain types of documents e.g. web pages referring to phone ring tones might be excluded so as not to be confused with music download sites with the phrase "ring tone" automatically causing a site to be rejected for further consideration.

SECOND EMBODIMENT

A second embodiment to the present invention will now be described with reference to Figures 6 and 7.

In the first embodiment a dedicated system for monitoring for unauthorised distribution of copyright works via the internet 3 was described. In contrast, in this embodiment a monitoring system 50 is provided which enables users to identify individual subjects and contexts that they wish to monitor via the internet 3.

The monitoring system 50 in accordance with this embodiment is identical to that described in the first embodiment except in place of the asset database 20 an input interface 52 is provided which is connected to the search context database 22, the search generation unit 24, the classification unit 30 and the classification database 32. The search generation unit 24 and classification unit 30 are also slightly modified as will be described in detail later.

In this embodiment the input interface 52 is arranged to enable a user to enter terms which are to be searched for on the internet 3 and to select contexts which the user wishes to identify the use of those terms. The search generation unit 24 then generates a set of seed terms for initiating a search of the internet 3 using the contexts and search term identified via the input interface 52. When web pages 11 are retrieved via the internet 3 by the search coordination unit 26, the classification unit 30 then proceeds to classify the retrieved web pages 11 using classification criteria associated with the contexts identified by the user input interface 52. In this way, the monitoring system 50 of this embodiment is able to utilise the search engines 9 on the search servers 5 to identify a large number of documents of

potential interest utilising the seed terms generated by the search generation unit 24. These web pages 11 of potential interest are then processed in detail by the classification unit 30 using the classification database 32 to filter identified pages against precise requirements.

Figure 7 is an illustration of an exemplary user interface for entering search terms and identifying context. When the monitoring system 50 of the present embodiment is first invoked, the input interface 52 generates a user interface such as that illustrated in Figure 7.

In this example, the user interface comprises a search term window 100, a number of context check boxes 101, a list of context names 102 adjacent to each of the context check boxes; an add context button 103; a modify context button 104; a search button 105 and a pointer 106.

Using conventional input devices such as a mouse or a keyboard a user then controls the pointer 106 to select the search term window 100, the context check boxes and any of the buttons 103-105. By selecting the search term window 100 and entering one or more search terms the user

identifies some seed terms which are to be utilised to generate a search.

By selecting the check boxes 101 adjacent to the list of contexts 102 a user identifies to the monitoring system 50 those contexts as identified by records within the search context database 22 and the classification database 32 against which a search using term identified in the search term window 100 is to be expanded using the search context database 22 and subsequently classified using the data within the classification database 32.

When terms have been entered within the search term window 100 and contexts selected from the list of contexts 102 a user then selects the search button 105 using the pointer 106 to cause the search terms within the search term window 100 to be passed to the search generation unit 24 and the search generation unit 24 to retrieve from the search context database 22 prestored data identifying how a search is to be expanded when contexts identified as being selected by the check boxes 101 are to be searched against. This data is then processed by the search generation unit 24 and the search coordination unit 26 as has previously been described in relation to the first embodiment.

When web pages 11 are retrieved by the search coordination unit 26 the classification unit 30 then classifies the retrieved data utilising classification rules within the classification database 32 associated with the contexts selected using the check boxes 101. Thus in this way by entering search terms and selecting contexts against which processing is to be made, the monitoring system 50 in this embodiment of the present invention is able to monitor for user specified search terms in defined contexts.

In this embodiment, in addition to the prestored data within the search context database 22 and the classification database 32 a user is able to modify or add additional context data against which a search may be made. In order for a user to add a new context to the list of contexts 102 a user initially selects the add context button 103 using the pointer 106. A user is then prompted to enter a name for the context, search expansion data for that context and rules for classifying documents for that context. The search expansion data terms are then stored within the search context database 22 and the rules for classifying the documents are stored within the classification database 32. When new context data has been entered this causes an additional context

check box 101 to appear on the user input interface together with the name of the newly input context. Subsequently, by selecting the new check box searches utilising the newly entered expansion terms and classification rules are then caused to be performed.

If a user wishes to modify either the list of search terms stored within the search context database 22 or the classification rules within the classification database 32 a user selects the modify context button 104 using the pointer 106 which enables a user to edit records stored within the search context database 22 and classification database 32. Thus in this way a user is able to tailor the manner in which the monitoring of web pages 11 by the monitoring system 50 is performed. Thus for example by entering a company name or trade mark or both within the search term window 100 and defining such expansion term and classification rules for context in which the trade mark is to be monitored against for example negative comments about a product which is to be monitored or general news comment about a particular company, the monitoring system 50 is able to identify the type of comment which is presently available via the internet 3 in relation to the search terms within the search term window in the identified contexts.

Further Modifications and Amendments

Whereas the invention has been described with reference to websites available via the Internet other sources of data could be used with the present invention. For example, databases available remotely could be interrogated periodically on the basis of search terms seeded by an agent as described herein. This could be of use with patents databases and publications databases of any nature. The results of those searches could be analysed, in the same way. In particular, each entry in a publications database normally includes an abstract of the publication, which could be utilised for relevance classification.

It will be appreciated that the present invention is not limited to monitoring only HTML websites. Other sources of information such as NNTP new groups could be monitored. Alternatively files in other file transfer protocols could be monitored.

In a system for monitoring NNTP news groups, initially the text of first page of a newsgroup might be analysed to determine whether the newsgroup was of interest. If it was determined to be the case, every page referred to on that front page could then be retrieved and the text

of each individual page analysed. Additionally, any hyperlinks within the pages could be extracted and the corresponding web pages analysed as has previously been described in detail.

Although the present invention has been described in terms of analysis of HTML data for web pages, similar analysis could be performed upon any form of data. For other types of data if it is possible to convert the data into HTML the data could be first converted and then analysis performed exactly as has previously been described. Alternatively processing of such data might be performed directly without any conversion of format. Thus for example an for FTP site storing music files suspect sites could be identified by analysing the file directory of the FTP site. If is is determined that the file names correspond to data identifying the product being monitored the FTP site could then be added to a list of suspected sites.

Monitoring of Internet resources of different types can also be performed simultaneously. Thus for example analysis of HTML web pages to identify links referring to FTP sites could be arranged to occur simultaneously with detailed analysis of directory file structures on

computers identified by the hyperlinks in the downloaded web pages.

Although in the first embodiment a monitoring system has been described where all web pages are classified prior to pages being selected for detailed analysis, it will be appreciated detailed analysis could take place immediately after a web page has been classified thereby avoiding the need for long term storage of the classification scores of web pages considered irrelevant.

The monitoring system 1 can be embodied by a plurality of computers, operable in parallel with separate processing power, and the search generation unit 24, the search coordination unit 26 and the classification unit 36 can be operable to allocate processes to be executed on different computers to manage processing resources effectively.

Whereas the present invention has been exemplified by a system for retrieval of information from "static" information sources such as websites it will be appreciated that the invention can also be applied to a system for retrieving information, processing that information and acting on the results of the processing.

For example, a system could be configured to retrieve stock market prices and other business information from particular sources and to perform calculations on the basis of that information to cause business transactions to be performed. These decisions can be configured in the rules language described herein, possibly with further decision making extensions to that language.

Further, a system in accordance with the invention could be configured to refer to websites offering shopping services, to compare prices and to give the user information concerning those prices so that the user can obtain the optimum price for goods or services which he may require.

Whereas the invention has been described in relation to specific example of searching websites it will be appreciated that any published source of information accessible via a computer network can be used in connection with the invention. For example, the system could be configured to monitor websites with rapidly changing content, such as those operated by newspapers or news gathering organisations, news groups, web bulletin boards which are similar to newsgroups but allow the posting of messages on a website handled in HTTP, and

chatrooms which provide a scrolling message recordal facility so that users can conduct conversations with other users.

The invention can also be applied to new protocols such as "hotline", Napster (for the exchange of audio information), ICQ (a messaging service), Gnutella and FastTrack.

It will be appreciated that, where the search coordination unit 26 is specified to schedule searches to be issued no more frequently than five seconds apart, the frequency of the schedule is capable of being altered to suit prevailing conditions. It may be the case that the administrator of a search engine may raise a complaint against the operator of the system of the illustrated example that search requests are being delivered thereto at too frequent a rate, in which case the search requests can be issued at a less frequent rate. Alternatively, the time period between search requests can be shortened in the event that it is perceived that searching is taking an unduly long time to be completed.

Whereas the system has been described in terms of a

computer network it will be appreciated that some elements of the user interface could be incorporated in an embedded system for implementation on a mobile communications device, such as a mobile telephone. In that way, a user would be able to make a query of a system in accordance with the present invention and to obtain collated results therefrom, or to obtain a simplified version of collated results therefrom. Such a system could take account of a limited communications speed between the mobile device and other devices, and limits the amount of data to be transferred accordingly.

Whereas the illustrated example is shown to demonstrate use of the present invention in discriminating and classifying words of the English language using word frequencies, it will be appreciated that similar techniques could be used to recognise other languages. In the case of agglomerative languages where words are frequently combined to produce longer, compound words, stemming may form a significant part of the language recognition process. Also, letter frequency, including analysis of the position of letters in words, could be used to recognise certain languages.

Certain languages are known to make little or no use of

particular letters, for example the letter "j" is very rarely used in Italian. Also, in the Italian language, many words end in a vowel. Each of these facts could be used to classify a document as being written in the Italian language.

Also, whereas the rules definition language described herein is expressed using words derived from English language words, it will be appreciated that other natural languages could be used as basis for the logical statements. Also, a more symbolic or graphical rule definition language could be used.

As an alternative or in addition to utilising text data to identify sites other forms of data could be used. Thus for example a corporate logo could be used to identify web sites relating to a company. For such a monitoring system, the presence of an image in a web site in a location relative to other discussion could cause a web page top be initially retrieved. The chance of selecting a web page for retrieval might be based upon the name given to the image file to be displayed. If a web page was selected for further analysis automatic comparison of a downloaded image and a copy of a particular logo could then be performed.

Thus for example monitoring websites for reference to Virgin Records could be performed where reference to an image such as virginlogo.bit caused a page to be analysed in great detail with the file virginlogo.bit being compared with a stored example bit image of the relevant corporate logo.

It will be appreciated that as in the case of the first embodiment by predicating further detailed analysis of a web page upon the web page may been determined to have at least some apparent relevance to a search being performed, the resources to perform detailed analysis such as pattern matching for specific images with images in web pages being monitored can be targeted and therefore performed more efficiently.

Appendix - Rule Language Description

Functions Reference

after(word1, word2)

Returns a true value if the first word appears after the second word in the document.

before(word1, word2)

Returns a true value if the first word appears before the second word in the document.

distance(word1, word2)

A value representing the number of words separating the two words identified as arguments of the distance function. Returns how far apart the two words are.

images()

Returns the number of images found in the document.

in_author(word)

Returns true if the word appears in the author identification section of the document.

in_bold(word)

Returns true if the word is in bold.

in_description(word)

Returns true if the word appears in the description of the document.

in_heading(word)

Returns true if the word appears in a heading.

in_heading1(word)

Returns true if the word appears in a heading style 1.

in_heading2(word)

Returns true if the word appears in a heading style 2.

in_heading3(word)

Returns true if the word appears in a heading style 3.

in_italic(word)

Returns true if the word is in italic.

in_keywords(word)

Returns true if the word appears in the keywords of the document.

in_link(word)

Returns true if the word appears in a link.

in_meta(word)

Returns true if the word appears in any meta-tag of the document.

in_title(word)

Returns true if the word appears in the title of the document.

in_underline(word)

Returns true if the word is in underline.

in_url(word)

Returns true if the word appears in the URL of the page.

links()

Returns the number of links found in the document.

near(word1, word2)

Returns true if the distance between the words is less than 20.

num_themes()

Returns the number of themes associated with the agent.

paragraph(word)

Returns the paragraph number of the word.

position(word)

Returns the word number of the word in the document.

print(string)

Outputs the string to the terminal.

printn(x)

Outputs the number x to the terminal.

section(word)

Returns the section number of the word.

sentence(word)

Returns the sentence number of the word.

sentence_position(word)

Returns the position of the word within a sentence.

sequence(word1, word2)

Returns true if the first word is immediately followed by the second word.

square(x)

Returns x*x.

words()

Returns the number of words in the document.

word_length()

Returns the average length of the words in the document.

CLAIMS:

1. Apparatus for identifying web pages for obtaining downloads of products, said apparatus comprising:

a data store operable to store items of data relating to a product;

a search unit operable to output instructions to search engines to perform searches and to retrieve web pages identified by said search engines; and

a selection unit operable to select from web pages retrieved by said search unit, web pages in which said items of data are identified as to be displayed in the vicinity of a hyperlink for downloading a digital file.

2. Apparatus according to claim 1 wherein said selection unit is operable to determine for hyperlinks within web pages, the type of file to be downloaded utilising said hyperlink and to select web pages in which said items of data are identified as to be displayed in the vicinity of a hyperlink for downloading a digital file of a predetermined type.

3. Apparatus according to claim 1 wherein said selection unit is operable to determine for hyperlinks within web pages, the size of files which said hyperlinks

are to be utilised to download and to select web pages in which said items of data are identified as to be displayed in the vicinity of a hyperlink for downloading a digital file of at least a predetermined size.

4. Apparatus according to claim 1 wherein said selection unit is operable to select from web pages retrieved by said search unit, web pages in which said items of data are identified as to be displayed in the vicinity of a hyperlink for downloading an audio file.

5. Apparatus according to claim 1 further comprising a classification unit operable to generate for web pages retrieved by said search unit one or more classification scores, wherein said selection unit is operable to select from web pages associated by said classification unit with predetermined classification scores.

6. Apparatus according to claim 5 wherein said classification unit is operable to generate said classification source on the basis of textual analysis of text data within said web pages.

7. Apparatus according to claim 1 wherein said search unit is operable to output instructions to search

engines, wherein said instructions are generated utilising items of data stored within said data store.

8. Computer apparatus for discriminating items of information comprising:

a search term store operable to store a search term for configuring a search engine to identify items of information for discrimination;

information retriever operable to retrieve items of information identified by a said search engine with respect to a said search term;

discrimination criterion store operable to store data defining a discrimination criterion to be applied to an item of information; and

information discrimination unit operable to apply said discrimination criterion to an item of information, to generate one or more classification scores for said item of information.

9. Information discrimination apparatus operable to receive and analyse items of information, comprising:

information receiver operable to receive an item of information for analysis;

one or more information analysis agents, the or each agent comprising at least one theme being an item of

textual information to be compared to said item of information for analysis, and one or more rules, the or each rule being a logical statement to be applied to said item of information for analysis; and

an analyser operable to compare a said theme with a said item of information to thereby generate a relevance score, and to apply a said rule or rules to a said item of information to thereby obtain one or more classification scores.

10. Apparatus according to claim 9 further comprising an instruction unit operable to send an instruction to a search engine at a remote location for items of information relating to a theme of one of said one or more information analysis agents, said information receiver being operable to receive an item of information on the basis of results of a search performed by said search engine.

11. Apparatus according to claim 10 wherein said instruction unit comprises search engine request scheduler means operable to manage configuration of one or more search engines to search in respect of themes.

12. Apparatus according to claim 10 wherein said

information receiver is operable to receive results information from a search engine, said results information comprising one or more identifiers to locations of items of information identified by said search engine as relevant to said search criterion, said information receiver comprising an identified item retrieval unit operable to retrieve the or each item of information from its location identified by said identifier.

13. Apparatus according to claim 12 further comprising an additional item identifier detection unit operable to detect if an identified item comprises an identifier to a location of an item of information, and operable to configure said identified item retrieval unit to retrieve from any detected identifier the corresponding item of information.

14. Apparatus according to claim 12 further comprising an identified item storage unit operable to store items of information retrieved by said identified item retrieval unit.

15. Apparatus according to claim 14 wherein said identified item retrieval unit is operable to compare a

retrieved identified item with information stored by said identified item storage unit, said identified item storage unit being operable to store said identified item on condition said identified item is not already stored by said identified item storage unit, items stored by said identified item storage unit in use having said analyser applied thereto.

16. Apparatus according to claim 9 wherein said information receiver comprises a retrieval schedule store operable to store a schedule for retrieval of items of information from identified remote locations, said information retrieval unit being operable to retrieve items of information in accordance with said schedule for analysis by said analyser.

17. Apparatus according to claim 9 wherein said analyser comprises a text data extractor operable to extract text data from a retrieved item of information, and wherein said analyser is operable to apply a text classification rule in one of said information analysis agents to text data extracted by said text extractor.

18. Apparatus according to claim 17 wherein said analyser comprises an image data extract or operable to

extract image data from a retrieved item of information, and wherein said analyser is operable to apply an image classification rule in one of said information analysis agents to image data extracted by said image extractor.

19. Apparatus according to claim 17 wherein said analyser comprises an audio data extractor operable to extract audio data from a retrieved item of information, and wherein said analyser is operable to apply an audio classification rule in one of said information analysis agents to audio data extracted by said audio extractor means.

20. Apparatus according to claim 17 wherein said analyser comprises video data extraction means for extracting video data from a retrieved item of information, and wherein said analysis means is operable to apply a video classification rule in one of said information analysis agents to video data extracted by said video extraction means.

21. Apparatus according to claim 17 wherein said analyser means is operable to apply a plurality of classification rules to a retrieved item of information and wherein said analyser comprises a classification

information collation unit for collating results of the application of said rules to said item of information.

22. Apparatus according to claim 21 wherein said analyser is operable to generate a numerical result in respect of application of a rule to an item of information, said collation unit being operable to collate numerical results into one or more cumulative totals.

23. Apparatus according to claim 17 wherein the or each information analysis agent stores the or each rule as text, and said analyser comprises a parser operable to parse said text to determine the classification rules.

24. Apparatus according to claim 23 wherein said parser is operable to identify a token in said data, said token being a keyword of a rule.

25. Apparatus according to claim 24 wherein said parser is operable to identify one or more tokens as arguments of an identified keyword token.

26. Apparatus according to claim 25 wherein said parser is operable to define a classification rule to which data

from a retrieved item of information can be applied.

27. Information discrimination apparatus operable to receive and analyse items of information, comprising:

information receiving means operable to receive an item of information for analysis;

information analysis agent storage means, for storing an information analysis agent, said storage means being operable to store; for an agent, a theme comprising an item of textual information for comparison with an item of information for analysis, and a rule comprising a logical statement to be applied to an item of information for analysis; and

analysis means for comparing a theme stored in said storage means with an item of information for analysis, and for applying a rule to said item of information, thereby to generate a relevance score with respect to said theme and a class score with respect to said rule for said item of information.

28. A method of discriminating items of information comprising:

storing a search term for configuring a search engine to identify items of information for discrimination;

retrieving items of information identified by said search engine with respect to said search term;

storing data defining a discrimination criterion to be applied to an item of information; and

applying said discrimination criterion to an item of information, to generate one or more classification scores for said item of information.

29. A method of analysing items of information comprising, on receipt of an item of information for analysis, the steps of:

comparing said item with an item of textual information defining a theme and from said comparison establishing a relevance score; and

applying to said item of information a logical statement being a rule resulting in generation of one or more classification scores for said information.

30. A method according to claim 29 comprising:

sending an instruction to an information source at a remote location for items of information relating to a theme and receiving an item of information on the basis of the results of a search performed by said search engine.

31. Method according to claim 30 wherein said method comprises storing a search criterion, and configuring a search engine at a remote location to search on the basis of a search criterion stored in said search criterion storing step.

32. Method according to claim 31 wherein said search engine configuring step comprises managing configuration of one or more search engines to search in respect of stored search criteria.

33. Method according to claim 31 wherein said receiving step comprises receiving results information from a search engine, said results information comprising one or more identifiers to locations of items of information identified by said search engine as relevant to said search criterion, and retrieving the or each item of information from its location identified by said identifier.

34. Method according to claim 33 further comprising detecting if an identified item comprises an identifier to a location of an item of information, and retrieving, in accordance with any detected identifier, the corresponding item of information.

35. Method according to claim 33 further comprising storing a retrieved item of information.

36. Method according to claim 35 comprising comparing a retrieved identified item with information stored in said storing step, and storing said identified item on condition said identified item is not already stored, and applying to items stored in said preceding step said stored discrimination data.

37. Method according to claim 29 wherein rule applying step comprises extracting text data from a retrieved item of information, and applying a text classification rule to text data extracted in said text data extracting step.

38. Method according to claim 29 wherein said rule applying step comprises extracting image data from a retrieved item of information, and applying an image classification rule to image data extracted in said image data extracting step.

39. Method according to claim 29 wherein said rule applying step comprises extracting audio data from a retrieved item of information, and applying an audio classification rule to audio data extracted by said audio

data extracting step.

40. Method according to claim 29 wherein said rule applying step comprises extracting video data from a retrieved item of information, and applying a video classification rule to video data extracted by said video data extracting step.

41. Method according to claim 37 wherein said rule applying step comprises applying a plurality of classification rules to a retrieved item of information and collating the results of the application of said rules to said item of information.

42. Method according to claim 41 wherein said rule applying step comprises generating a numerical result in respect of application of a rule to an item of information, said collating step comprising collating numerical results into one or more cumulative totals.

43. A method according to claim 37 wherein said rule applying step comprises parsing data stored to define said rule into a logical statement to define a classification rule.

44. A method according to claim 43 wherein said parsing step comprises identifying a token in said data as a keyword of a rule.

45. A method according to claim 44 wherein said parsing step comprises identifying one or more arguments of an identified keyword token.

46. A method according to claim 45 wherein said parsing step comprises defining a classification rule on the basis of an identified keyword and one or more identified arguments, to which data from a retrieved item of information can be applied.

47. A computer program product comprising processor executable instructions operable to configure a computer to become operable as apparatus in accordance with any of claims 1 to 27.

48. A computer program product comprising processor executable instructions operable to configure a computer to perform a method in accordance with any of claims 28 to 46.

49. A system comprising a computer apparatus in

accordance with any of claims 1 to 30 and a user terminal, said user terminal comprising:

user instruction receiving means for receiving a user instruction for initiating operation of said computer apparatus for retrieving and discriminating items of information;

discrimination information receiving means for receiving, from said retrieving and discriminating apparatus, discrimination information identifying items of information including one or more themes and one or more rules; and

output means for outputting said information to a user.

50. A user terminal for use in a system according to claim 49, comprising:

user instruction receiving means for receiving a user instruction for initiating operation of said computer apparatus for retrieving and discriminating items of information;

discrimination information receiving means for receiving, from said retrieving and discriminating apparatus, discrimination information identifying items of information; and

output means for outputting said information to a

user.

51. A computer program product comprising processor executable instructions operable to configure a computer as a user terminal in accordance with claim 50.

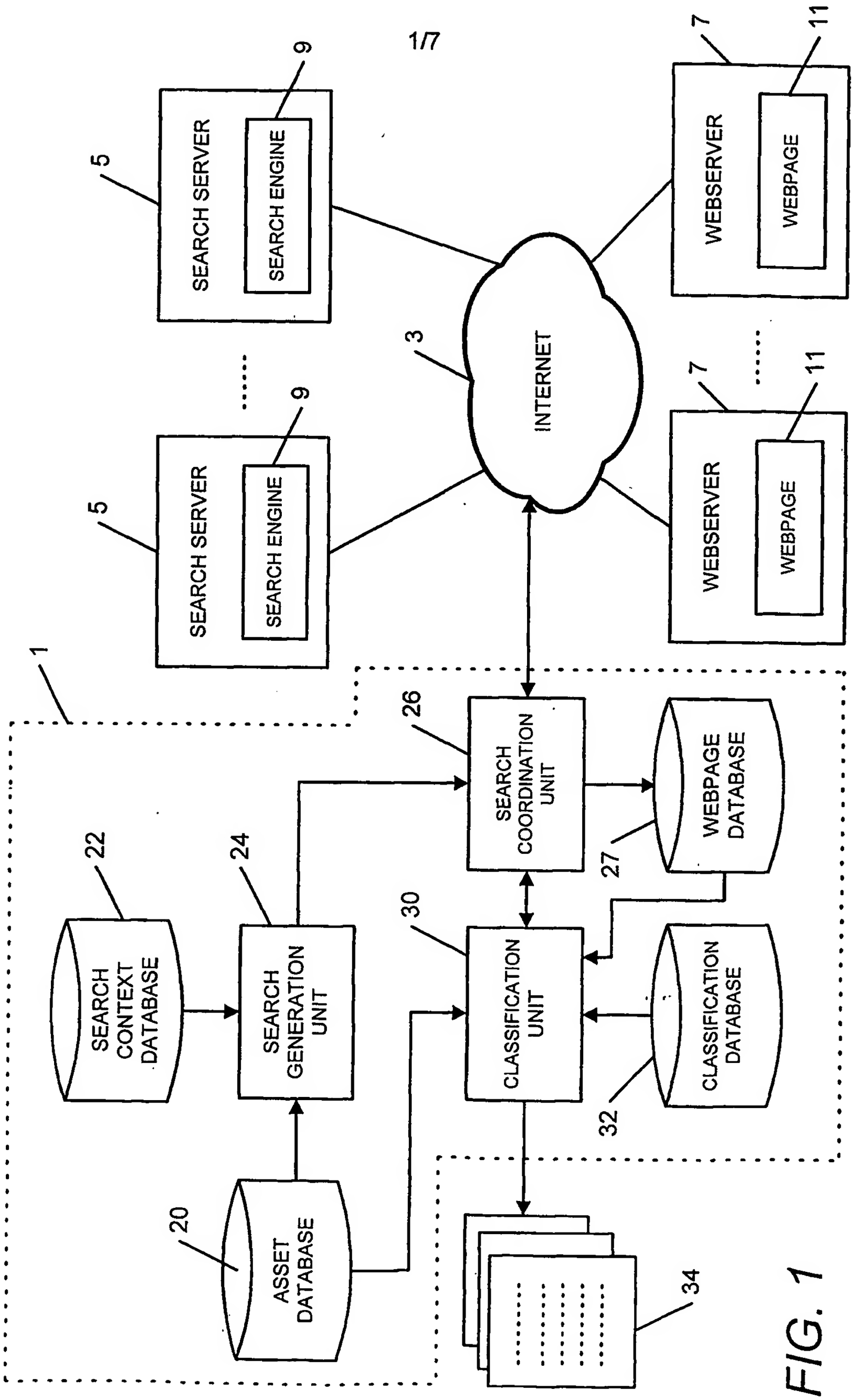

SEARCH SERVER
SEARCH ENGINE
5
9
INTERNET
3
WEBSERVER
WEBPAGE
7
11
1
SEARCH COORDINATION UNIT
26
WEBPAGE DATABASE
27
SEARCH CONTEXT DATABASE
22
SEARCH GENERATION UNIT
24
CLASSIFICATION UNIT
30
CLASSIFICATION DATABASE
32
ASSET DATABASE
20
34

FIG. 1

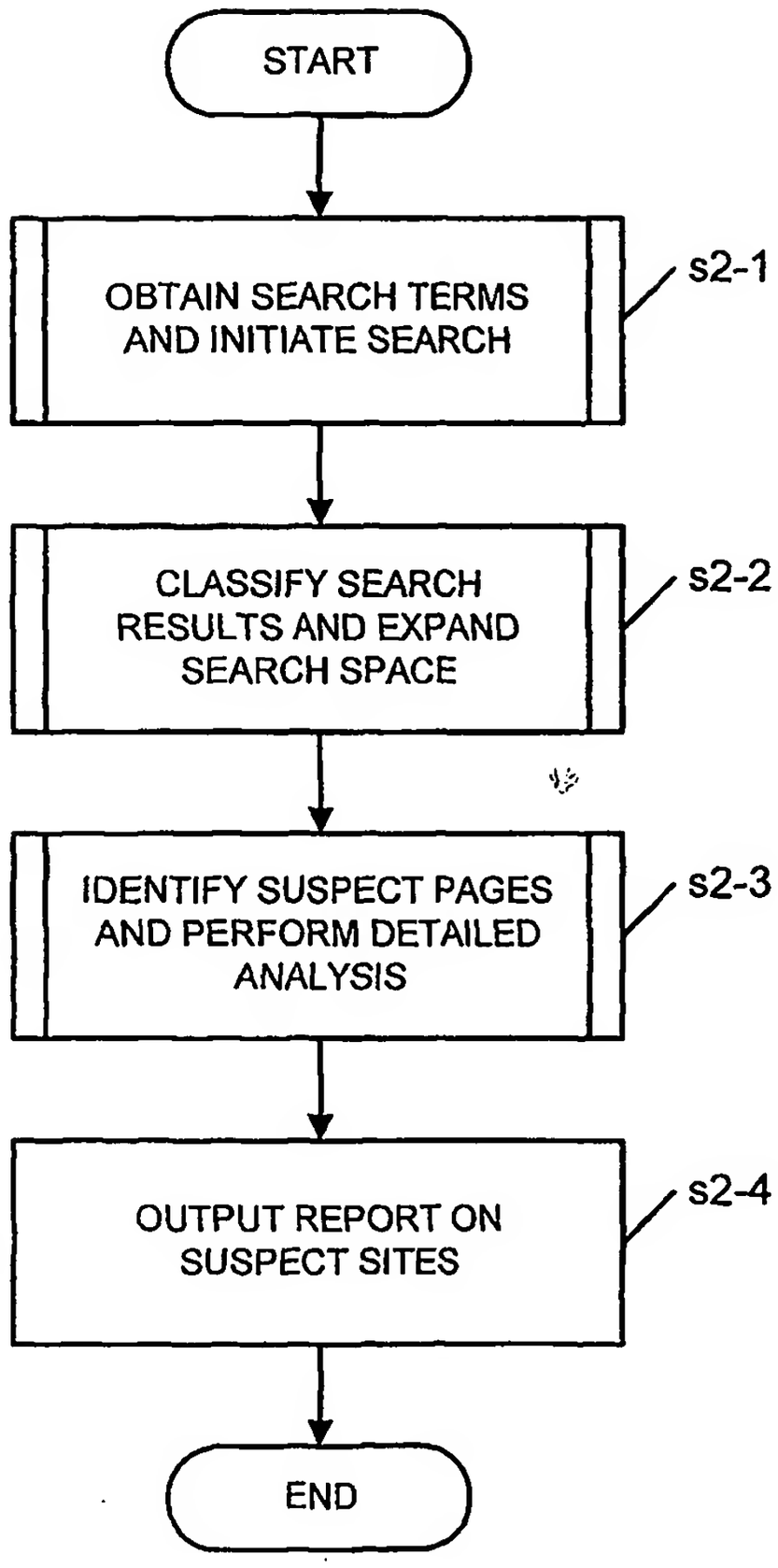

START
OBTAIN SEARCH TERMS AND INITIATE SEARCH
s2-1
CLASSIFY SEARCH RESULTS AND EXPAND SEARCH SPACE
s2-2
IDENTIFY SUSPECT PAGES AND PERFORM DETAILED ANALYSIS
s2-3
OUTPUT REPORT ON SUSPECT SITES
s2-4
END

FIG. 2

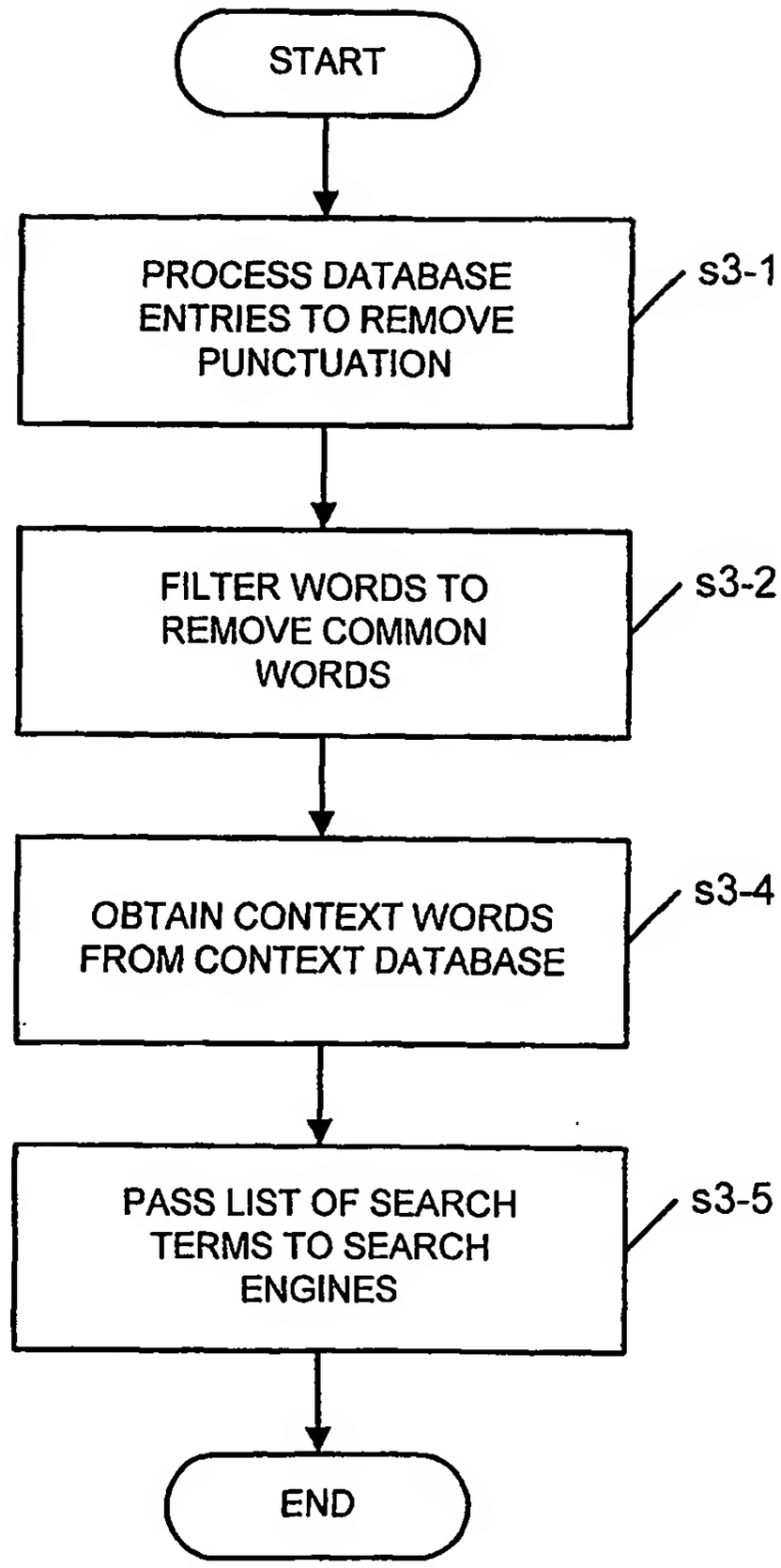
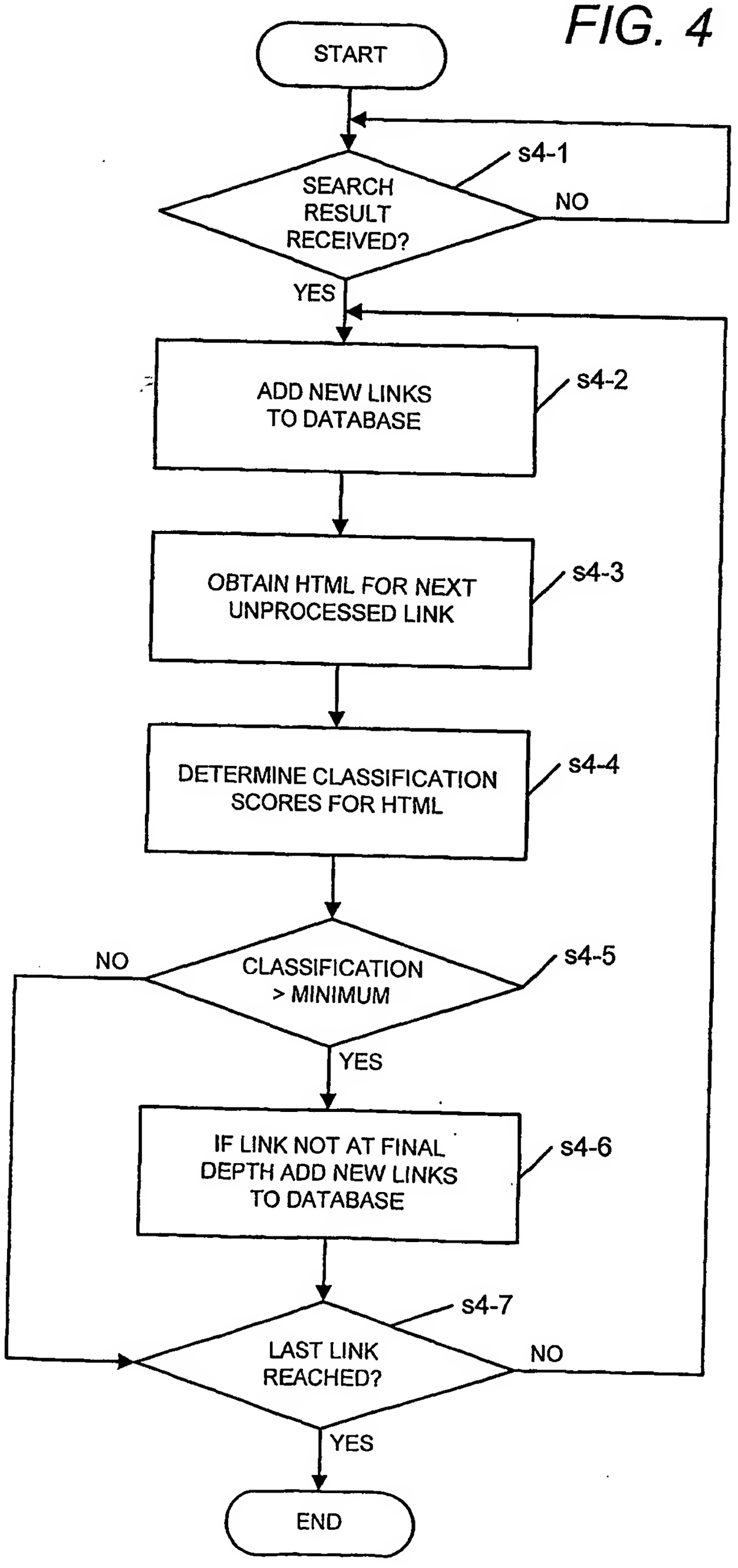

START
PROCESS DATABASE ENTRIES TO REMOVE PUNCTUATION
s3-1
FILTER WORDS TO REMOVE COMMON WORDS
s3-2
OBTAIN CONTEXT WORDS FROM CONTEXT DATABASE
s3-4
PASS LIST OF SEARCH TERMS TO SEARCH ENGINES
s3-5
END

FIG. 3

FIG. 4


START
s4-1
SEARCH RESULT RECEIVED?
NO
YES
ADD NEW LINKS TO DATABASE
s4-2
OBTAIN HTML FOR NEXT UNPROCESSED LINK
s4-3
DETERMINE CLASSIFICATION SCORES FOR HTML
s4-4
NO
CLASSIFICATION > MINIMUM
s4-5
YES
IF LINK NOT AT FINAL DEPTH ADD NEW LINKS TO DATABASE
s4-6
s4-7
LAST LINK REACHED?
NO
YES
END

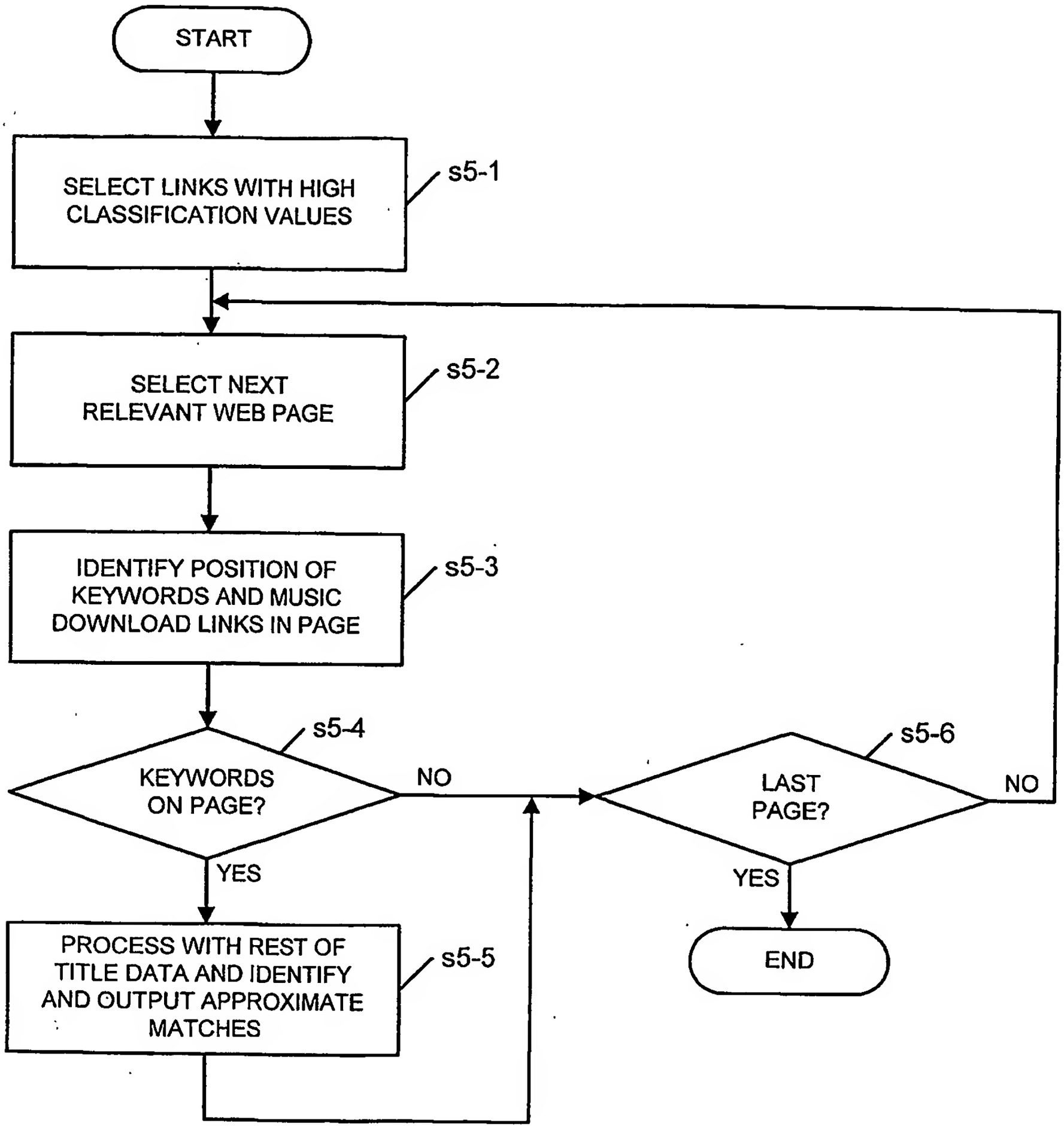

START
SELECT LINKS WITH HIGH CLASSIFICATION VALUES
s5-1
SELECT NEXT RELEVANT WEB PAGE
s5-2
IDENTIFY POSITION OF KEYWORDS AND MUSIC DOWNLOAD LINKS IN PAGE
s5-3
KEYWORDS ON PAGE?
s5-4
NO
YES
PROCESS WITH REST OF TITLE DATA AND IDENTIFY AND OUTPUT APPROXIMATE MATCHES
s5-5
LAST PAGE?
s5-6
NO
YES
END

FIG. 5

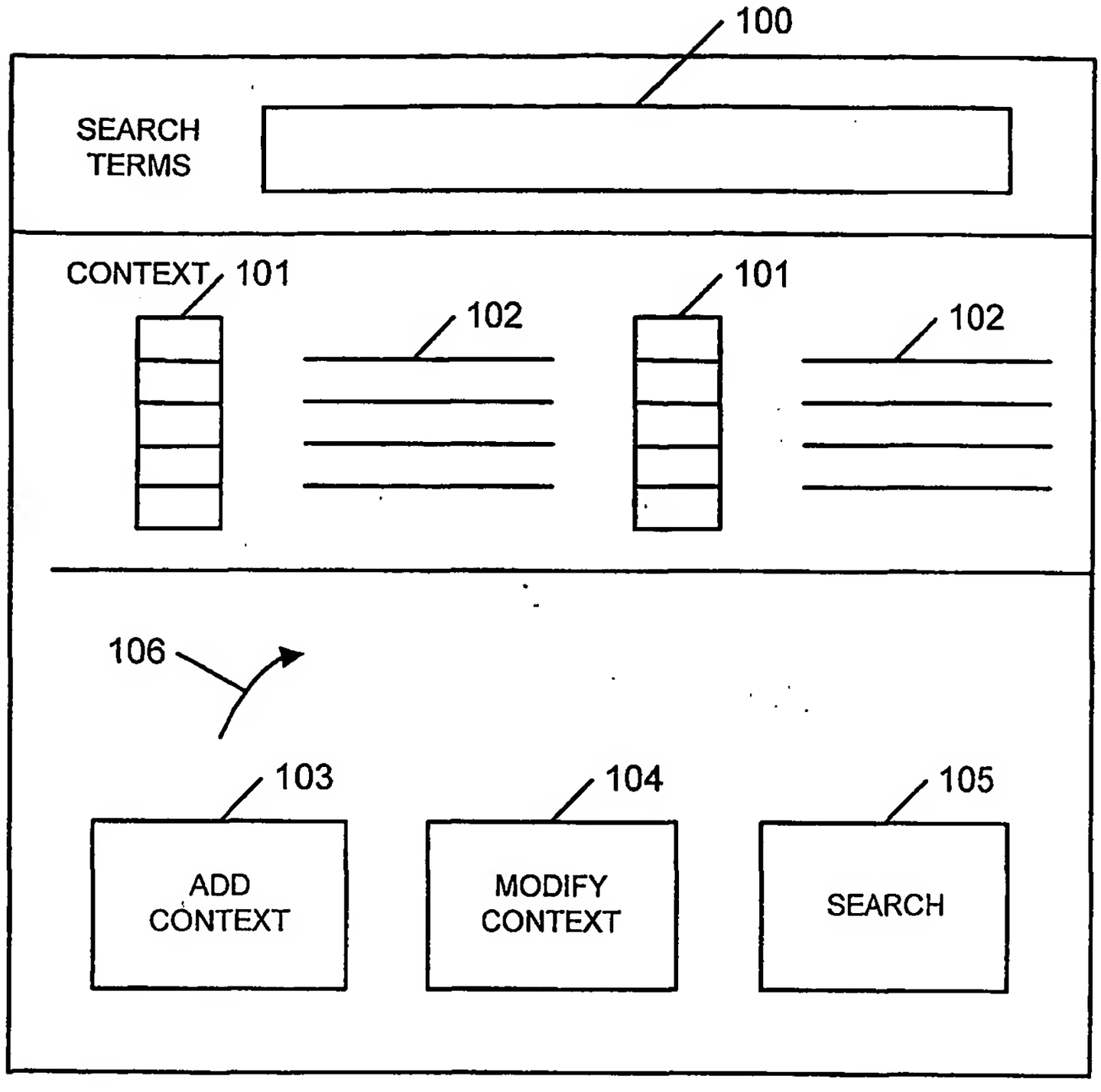

100
SEARCH TERMS
CONTEXT
101
102
101
102
106
103
ADD CONTEXT
104
MODIFY CONTEXT
105
SEARCH

FIG. 7

INTERNATIONAL SEARCH REPORT

In tional Application No
PCT/GB 01/04869

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 822 502 A (BRITISH TELECOMM) 4 February 1998 (1998-02-04) page 3, line 28 -page 4, line 25 page 5, line 41 -page 6, line 15 page 6, line 45 -page 7, line 39 ---	1-51
X	LAWRENCE S ET AL: "Inquirus, the NECI meta search engine" COMPUTER NETWORKS AND ISDN SYSTEMS, NORTH HOLLAND PUBLISHING. AMSTERDAM, NL, vol. 30, no. 1-7, 1 April 1998 (1998-04-01), pages 95-105, XP004121436 ISSN: 0169-7552 abstract page 97, left-hand column, line 6 -page 98, left-hand column, line 28 page 98, right-hand column, line 21-27 --- -/--	1-51

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 April 2002	22/04/2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Correia Martins, F

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Ir ational Application No
PCT/GB 01/04869

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	US 6 012 053 A (SCHIEGG MICHAEL J ET AL) 4 January 2000 (2000-01-04) column 1, line 18-31 column 2, line 4-24 column 3, line 33-63 column 6, line 34-65 column 7, line 23-30 ---	1-51
X	WO 99 12108 A (WEEKS RICHARD ;BRITISH TELECOMM (GB); DAVIES NICHOLAS JOHN (GB)) 11 March 1999 (1999-03-11) page 5, line 5-14 page 7, line 26 -page 10, line 22 page 15, line 23-26 -----	1-51

INTERNATIONAL SEARCH REPORT

Information on patent family members

In ational Application No
PCT/GB 01/04869

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
EP 0822502	A	04-02-1998	EP	0822502 A1	04-02-1998
			AU	3554297 A	20-02-1998
			EP	0917683 A1	26-05-1999
			WO	9804979 A1	05-02-1998
			JP	2000516005 T	28-11-2000
			US	6178419 B1	23-01-2001
US 6012053	A	04-01-2000	NONE		
WO 9912108	A	11-03-1999	AU	742831 B2	10-01-2002
			AU	8876298 A	22-03-1999
			CA	2302264 A1	11-03-1999
			CN	1269897 T	11-10-2000
			EP	1010105 A1	21-06-2000
			WO	9912108 A1	11-03-1999
			JP	2001515245 T	18-09-2001
			US	6353827 B1	05-03-2002